Convolution Neural Network in Visual Tracking using Correlation Filter

V. ISWARYA [1], D.SINDHU [2]
[1,2] Dept. of Computer Engineering, Dr. Sivanthi Aditanar College of Engineering, Tiruchendur

Abstract -- In the past years correlation filter have shown impressive results for visual object tracking. The types of features present in this tracker affect the performance of visual object tracking. The goal is to utilize object detection features whenever change the appearance of the object. In this project correlation filter is invoked in convolution neural network and find a location of object. Correlation filter based (CFB) trackers it used the network for classification problem. Based on the loss function of network a back propagation algorithm is used for the proposed model. The newly proposed model is flexible with custom design and also makes dependency on network trained for classification. Convolution part of state of art network must be fine-tuned and get achieved in performance by 20%.Tracking failures must be decreased by 30% when we use the tracking dataset VOT-2016(visual object tracking)

Indexed Terms: visual tracking, correlation filter, correlation plane, Loss function

I. INTRODUCTION

Visual Tracking is a fundamental task in computer vision which has been extensively researched. Though much progress exists in literature, it is still very challenging due to factors such as partial occlusions, pose variations, viewpoint variations and so on. In visual tracking we mark every object at the beginning of video sequence to identify the target object. Tracking is done by predicting the object at each frame. The benchmark dataset used to predict the performance of the tracking algorithms by applying a ground truth bounding box must be done in every object. The various machine learning concept Such as sparse generative methods, support vector machines and deep learning must be introduced for improving the tracking performance. In this project we use a convolution neural network for image classification. A convolution neural network (CNN or Conv Net) is one of the most popular algorithms for deep learning, a type of machine learning in which a model learns to perform classification tasks directly from images, video, text, or sound. CNNs are particularly useful for finding patterns in images to recognize objects, faces, and scenes. They learn directly from image data, using patterns to classify images and eliminating the need for manual feature extraction.

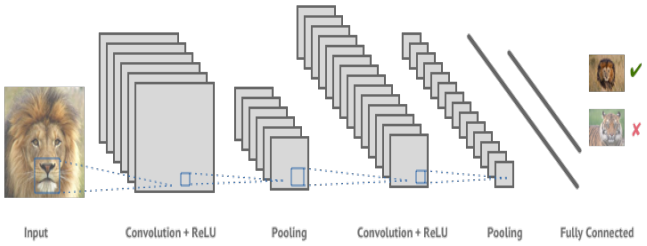

Fig. 1: Example of a network with many convolution layers

The ultimate goal of the project is to utilize object detection feature maps for visual tracking. Filters are applied to each training image at different resolution and output of one image must be convolved as input of next image. We have to train the dataset into correlation filter based tracking methods such as DSST (Discriminative Scale Space Tracker) and CCOT (Continuous Convolution Operator Tracker) it mainly increase the performance of benchmark datasets such as VOT-2015, VOT-2016 (Visual Object Tracking) and OTB-2013, OTB-2015

II. RELATED WORK

The various methods have been introduced for solving the problem of visual object tracking. NowIn this section we provide a survey and also provide a link between proposed methods and previous method.

A. Discriminative Trackers:

Discriminative methods use a classifier model for classification of object or background. Training is performed by gather information from their interest

when they begin tracking. The location of object is done by seeing a candidate location score must be high. In this method at each candidate location the Classifier must be evaluated.

B. Generative Methods:

It mainly describes the appearance of the object. According to the test instance the object location is estimated. From the predicted location obtained the model is obtained using a test instance the method proposed is online subspace learning method. The other method from the object appeared in brightness histogram of object patch. Non negative matrix factorization is also form a visual tracking problem.

C. Correlation Filter Based Trackers:

Correlation filter is used to minimize the squared error between the desired plane and the correlation plane. The correlation filter cost is finding in frequency domain using the convolution theorem and fast Fourier transform. It also extends the use of multi- channel support for increase the tracking Performance Kernelized correlation filter (KCF) are used spatial regularization for imperfect training example. Pre-trained CNN are used the feature map at different resolution. The main drawback is training a small network model.

D. Custom Architecture for Visual Track:

The Siamese feature learning method applied to visual tracking it mainly represent the output with similar features for target object and dissimilar for target and non-target samples. But in this learning method evaluation of candidate must be expensive. Then the convolution neural network must be introduced it directly find the target object location. The test frames are passed to the same convolution layers using the fully connected neural network. We use the recurrent neural network it mainly used to find the confidence map of target object. Spatial relationship is used between the object and the background.

E. Combining Trackers:

Multiple online tracker must be combined for visual tracking .Object detection must be accomplished using the part based version of MOSSE. The trackers must be sample and combine using a Markov Chain Monte Carlo sampling method. The generative and discriminative approaches must be combined using the hybrid methods. When we suffer from the heavy computation load discriminative network use a tree structure it mainly store different nodes of tree in CNN models.

III. PROPOSED FRAMEWORK

The two correlation filter based tracking methods used in this visual tracking. They are Discriminative Scale Space Tracker (DSST) and Continuous Convolution Operator Tracker (CCOT). From the learned features the framework must split in to trackers used in the benchmark datasets.

- The Proposed framework consist of single fully convolution neural network
- Training of this model is performed by propagating two image patches which contain the same visual object, through the model
- Once the feature maps are obtained for each image patch, the correlation filter is calculated from the template patch.
- The reduction of the difference between this two signals is obtained by the back propagation of the error and the stochastic gradient descent procedure

Object tracking is the process of locating and moving objects or multiple objects over the time in video. An output of object tracking is in object track. It is the sequence of object location in each frame of video.

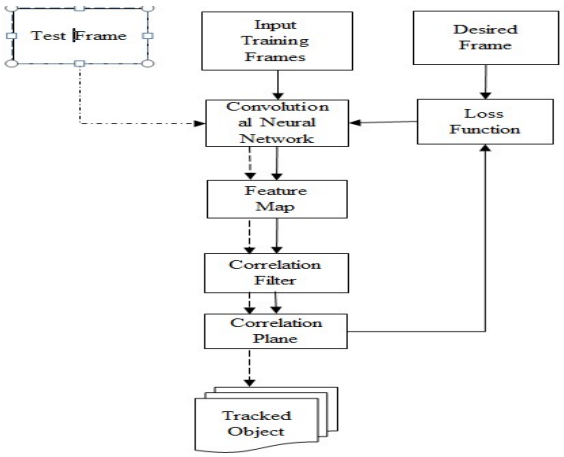

Fig. 2: Method Overview: A CNN model is used to track the object using the correlation filter in visual tracking.

A. INPUT TRAINING FRAME:

The dataset can be presented in various forms. Some of them are as follows they are

- Collection of sequence of consecutive or related frames as a single input item.
- Collection of frames in random order
- Collection of frames in sequential order.

For training a network you should have a label for input data. The network will be self-revise using that loss value by back propagating. So this process is called as training. We have an input data not label so loss cannot be calculated. So fine tuning is almost same word with 'additional training so you cannot fine train your retrained network without labelled data.

B. CONVOLUTIONAL NEURAL NETWORK:

Convolution neural network are very similar to ordinary neural networks. A convolution neural network consists of an input layer and output layer as well as multiple hidden layers.

- A convolution neural network (CNN) is one of the most popular algorithms for deep learning, a type of machine learning in which a model learns to perform classification tasks directly from images, video, text, or sound
- CNNs are particularly useful for finding patterns in images to recognize objects, faces, and scenes. They learn directly from image data, using patterns to classify images and eliminating the need for manual feature extraction

The hidden layers of a convolution neural network consist of layers such as

- Convolution layers
- Pooling layers
- Fully connected layers

C. CONVOLUTION LAYERS:

A convolution neural network is used for the object recognition and computer vision and also object recognition.

- Convolution neural network eliminate the need for manual feature extraction- the features are learned directly by convolution neural network
- Convolution neural network produce state of the art recognize results
- Convolution neural network can be retrained for new recognition tasks enabling us to build on pre-existing networks

Convolution puts the input images through a set of convolution filters each of which activates certain features from the images.

D. POOLING LAYERS:

Pooling layers simplifies the output by performing nonlinear down sampling reducing the number of parameters that the network needs to learn.

E. RECTIFIED LINEAR UNIT (RELU):

Rectified linear unit allows for faster and more effective training by mapping negative values to zero and maintaining positive values. This is sometimes referred to as activation because only the activated features are carried forward into next layer.

F. FEATURE MAP:

The feature map is the output of one filter applied to the previous layer .Each position result in the activation of the activation of neuron and the output is collected in the feature maps. The feature map is further processed by the correlation filter.

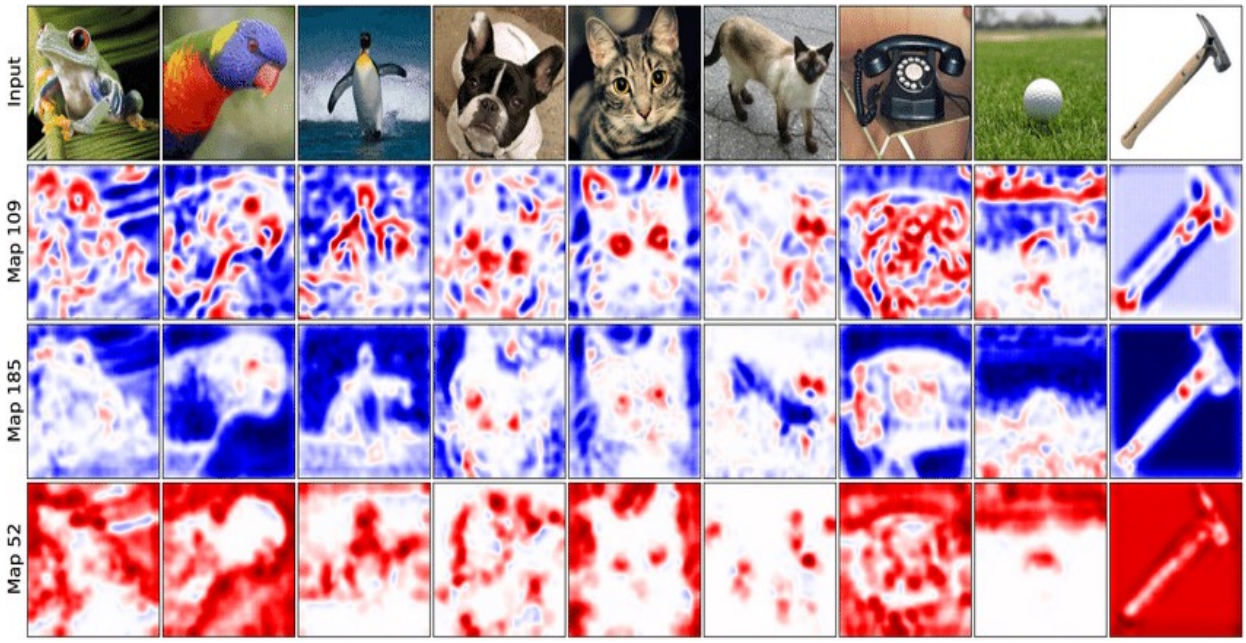

Fig. 3: Feature Map

G. CORRELATION FILTER

The Correlation filters are a class of classifiers which are specially optimized to produce the sharp peaks in the correlation output primarily to achieve accurate localization of targets in scenes. First traditional correlation filter designs are limited to scalar feature representation of objects.

- The correlation filter is the algorithm that trains a linear template to discriminate between the images and their translations
- It is mainly well suited to object tracking
- Its formulation in the Fourier domain provides a fast solution enabling the detector to be re-trained per once frame

One can view a correlation filter in the spatial domain as linear least scores discriminate. Made popular by Bole et.al; referred to in literature as a Minimum Output Sum of Squared Error (MOSOSE)

H. CORRLEATION PLANE:

The object with highest matching score is considered as tracked target. A fast normalized cross correlation is used to match template with the object in every frame of video.

- In two dimensions. In the real projective plane, points and lines are dual to each other
- A correlation is a point-to-line and a line-to-point transformation that preserves the relation of incidence in accordance with the principle of duality.
- The correlation plane also estimate the high correlation in target object

I. LOSS FUNCTION:

Loss function is used to measure the degree of fit. A loss function Loss(x, y, w) quantifies how unhappy you would be if you used w to make a prediction on x when the correct output is y. It is the object we want to minimize. By using this measure BPN is updated the weight values.

- In a neural network, this is done using back propagation
- The current error is typically propagated backwards to a previous layer, where it is used to modify the weights and bias in such a way that the error is minimized.
- The weights are modified using a function called optimized function.

J. COMPUTATIONAL COMPLEXITY:

The gradient terms must be calculated in DFT domain with all the element-wise multiplication, division summation and DFT transform operation .The complexity of the DFT calculation is given by O

(P log (p)) where P is the length of the signal. The classification networks such as Alex net and VGG an auxiliary layer with fewer feature maps are added on the top of convolution layers. It is mainly found that localization improves as the feature maps must be increased. The amount of quality improvement reduces as the distance between the layer increases.

IV. EXPERIMENTAL RESULTS

A. DATASET GENERATION:

The proposed tracker configurations are found on OTB-2013, OTB-2015 and VOT-2015. VOT-2016 is the 2016 most challenge called visual object tracking. The performance must be evaluated between two metrics they are (1) success curve (2) precision curve

B. DISCRIMINATIVE SCALE SPACE TRACKER:

Robust scale estimation is a challenging problem in visual object tracking. The high correlation object will be tracked and we find the output from the bounding box. The bounding box will be estimated from seeing the ground truth of the object in visual tracking methods.

C. OUTPUT:

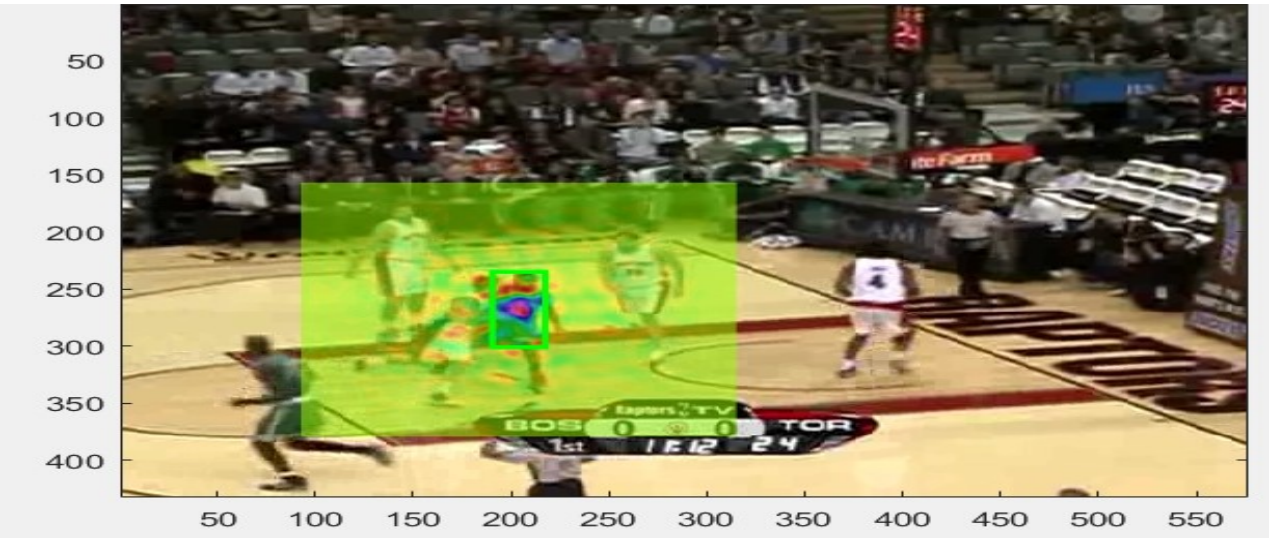

Fig. 4: Tracked the Person in every frame

D. PERFOMANCE EVALUATION:

We analyse the performance how the proposed learning features framework at different training configuration must be used. The performance of using correlation plane found the high values of object tracking. Loss function can be predicted to find the object track at high proportions. Hence the performance of convolution neural network must be satisfied.

V. CONCLUSION

By exploiting the correlation theorem, an efficient back propagation formulation is presented. The introduced feature learning method is trained on the frames generated by utilizing VOT2015 and ILSVRC Video datasets. Thus the feature learning problem for correlation filter based visual tracking task is resolved. The back propagation technique must be used to train convolution neural network using stochastic gradient descent.

REFERENCES

[1] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark." in CVPR. IEEE, 2013, pp. 2411–2418. [Online].

[2] "Object tracking benchmark." IEEE Trans. Pattern Anal. Mach. Intel., vol. 37, no. 9, pp. 1834–1848, 2015.

[3] C. Bao , Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in CVPR, IEEE Conference on, June 2012, pp. 1830–1837.

[4] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski , "Robust and fast collaborative tracking with two stage sparse optimization," in IEEE ECCV, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6314, pp. 624–637. [Online].

[5] S. Hare, A. Saffari, and P. Torr, "Struck: Structured output tracking with kernels," in ICCV, Nov 2011, pp. 263–270.

[6] H. Nam and B. Han, "Learning multi-domain convolution neural networks for visual tracking," in The IEEE Conference on CVPR, June 2016.

[7] H. Li, Y. Li, and F. Porikli, "Deep track: Learning discriminative feature representations by convolution neural networks for visual tracking," in Proceedings of the BMVC.BMVA Press, 2014.